SMOTE 数据预处理算法在砂型铸造复杂铸件缺陷预测中的应用

潘徐政[1]，刘迎辉[1]，李　文[1]，计效园[1]，殷亚军[1]，吴来发[2]，解明国[2]，周建新[1]

（1. 华中科技大学材料成形与模具技术国家重点实验室，湖北武汉 430074；
2.安徽合力股份有限公司，安徽合肥 230601）

摘要：针对实际生产过程采集的复杂转向桥铸件工艺数据中冷隔、气孔、砂眼、缩孔等缺陷类别的数据量严重不平衡、复杂铸件缺陷预测模型准确率不高的问题，结合砂型铸造实际工况，引入了SMOTE（Synthetic Minority Oversampling Technique）数据预处理算法，探究其在砂型铸造复杂铸件缺陷预测中的应用。根据采集到的复杂铸件不平衡数据集的特点，基于SMOTE数据预处理算法，科学扩充了不平衡数据集，创建了可用于训练复杂铸件缺陷预测模型的平衡数据集，数据预处理前后的模型预测准确率从86.50%提高至97.91%。

关键词：转向桥铸件；砂型铸造；不平衡数据集；数据预处理；SMOTE算法；缺陷预测

作者简介：
潘徐政（1999–），男，博士生，研究方向为铸造数字化管理系统。E-mail：2028119722@qq.com
通讯作者：
计效园，男，副教授。E-mail：jixiaoyuan@hust.edu.cn

中图分类号：TP391.3
文献标识码：A
文章编号：1001-4977（2024）10-1473-07

基金项目：
国家重点研发计划项目（2020YFB1710100）；国家自然科学基金（52275337、52090042、51905188）。
收稿日期：
2023–09–13 收到初稿，
2023–12–15 收到修订稿。

砂型铸造是铸造业的重要组成部分，复杂铸件在砂型铸造生产过程中，由于工艺参数多且生产环境复杂，铸件常会出现各类缺陷。数据驱动的复杂铸件缺陷预测等人工智能方法能够有效利用预测结果指导生产过程，有利于降低铸件产生缺陷的概率，帮助企业提高生产效益。我国是铸造大国，复杂铸件的生产数据量足够满足数据驱动的要求。但是对复杂铸件生产数据进行采集，得到的绝大部分是无缺陷铸件的生产数据，各类缺陷的生产数据量相比于合格铸件的生产数据量远远不够，存在数据不平衡问题。为解决数据不平衡问题，需要引入数据预处理的方法。

2020年，东北大学徐赵兆等[1]针对基于C4.5决策树算法模型对不平衡医疗数据的分类精度不理想的问题，提出了一种结合合成小波过采样技术的基于聚类的Oversampling算法和K-means算法。2020年，香港大学硕峰等[2]针对基于SMOTE的过采样技术性能高度不稳定的问题，提出了一系列基于SMOTE的稳定过采样技术。2021年，西北工业大学刘少伟等[3]针对实际工程中收集的滚动轴承故障数据严重不平衡问题，提出了一种深度特征增强生成对抗网络的新型数据预处理方法。2020年，印度圣彼得高等教育研究所Ulagapriya Krishnan、Pushpa Sangar[4]通过在机器学习中应用不同的采样技术，提高不平衡数据的分类性。2020年，重庆大学段梦明等[5]针对移动系统的数据分布不平衡问题，证明了不平衡的分布式训练数据将导致联合学习应用的准确性降低，并建立了一个名为Astraea的自平衡框架。2021年，武汉大学荆晓远等[6]针对现有的不平衡学习方法在分类性能方面精度不足的问题，提出了一种成本敏感的多网学习（UCML）方法。2020年，北京交通大学王新越等[7]针对过采样生成实例应该生成多少合成实例的问题，提出了一种基于局部分配的自适应少数群体过采样方法。2016年，美国普雷斯克岛缅因州大学Mujahid N. Syed等[8]针对气压系统数据高阶不平衡问题，提出了一种用于气压系统故障检测的新机器学习方法。2021年，巴基斯坦伊斯兰堡大学Yousra Asim等[9]针对基于博主和博客特征的非结构化数据和不平衡数据问题，提出基于案例的推理、影响博客预测的新颖框架。2017年，英国剑桥大学Yasmin Fathy等[10]针对不平衡数据限制机器学习预测断层的问题，提出了一个全面的比较分析应用建议框架，分析了算法组件不同组合的性能，深入研究了

应用程序上下文和机器学习算法设计之间的相关性。

本研究针对在实际生产过程中采集到的复杂转向桥铸件工艺数据严重不平衡的问题，引入SMOTE数据预处理算法，科学扩充了不平衡数据，并对比了数据预处理前后缺陷预测模型的性能，得到较优的预测结果。

1 数据获取与数据预处理

1.1 数据获取

本研究的数据来源为工程机械叉车转向桥铸件的生产数据，转向桥铸件材质为球墨铸铁，牌号为QT450-10，承担转向任务，具有复杂结构，除了承担汽车的垂直载荷外，还承受纵向力和侧向力及这些力造成的力矩作用，对铸件强度、抗疲劳性能等力学性能有极高的要求，是工程机械车辆中的关键传动装置[11]，转向桥铸件三维图如图1所示。

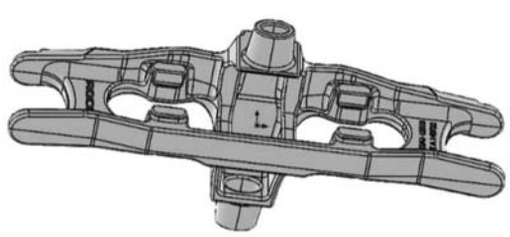


图1 转向桥铸件三维图

Fig.1 3D drawing of steering bridge casting

从企业获取的生产数据包含无缺陷数据5 848条，冷隔缺陷数据274条，气孔缺陷数据399条，砂眼缺陷数据359条，缩孔缺陷数据148条。

1.2 数据预处理

原始数据不仅包含与铸件生产质量密切相关的记录，如化学成分、浇注参数、砂处理参数等关键参数，还包含与铸件生产质量无关的记录。经过数据预处理，将无关记录去除之后，数据样式如表1所示。

2 数据不平衡问题与解决方法

2.1 数据不平衡问题

大多数分类学习方法都有一个共同的基本假设，即不同类别的训练样本数量相等。如果不同类型的训练样本数量有细微的差异，通常影响不大。但是如果差异过大，就会给学习过程带来麻烦。

例如，如果反例有900个，而正例只有100个，则该学习方法只需要返回一个总是预测新样本为反例的学习器，准确率就可以达到90%。但是根据这样的训练样本训练出的学习器是没有价值的，因为它不能预测任何正例。

经过简单预处理得到的数据集，有约6 000条无缺陷数据，仅有约1 000条有缺陷数据；若把缺陷分类，则每种缺陷拥有的数据会更少。这是典型的不平衡数据，利用不平衡数据训练出的预测模型无法作出准确的预测。

表1 企业铸件生产数据样式表

Table 1 Enterprise casting production data style table

合金	化学成分w_B/%							浇注参数			孕育率/%	砂处理参数							缺陷
	C	Si	Mn	P	S	Mg	Al	浇注重量/kg	浇注温度/℃	浇注时间/s		紧实率/%	剪切强度/MPa	旧砂温度/℃	旧砂水分/%	膨润土/%	混配土/%	新砂/%	不良原因
转向桥-01	3.71	2.68	0.62	0.022	0.010	0.042	0.022	138.2	1 407	15.2	30	36.88	5.07	33.4	1.78	23.9	11.4	0	无
转向桥-02	3.85	2.70	0.52	0.020	0.009	0.042	0.027	134.1	1 388	15.9	61	40.06	5.51	39.2	2.19	24.4	11.4	26	冷隔
转向桥-03	3.74	2.68	0.56	0.022	0.010	0.040	0.023	133.8	1 410	20.5	76	41.41	5.38	36.5	2.00	21.7	12.1	15	气孔
转向桥-04	3.68	2.82	0.51	0.022	0.011	0.043	0.026	139.7	1 399	15.6	39	47.43	2.00	37.6	2.11	21.5	12.3	0	砂眼
转向桥-05	3.72	2.69	0.52	0.033	0.014	0.054	0.024	132.0	1 406	16.1	51	40.9	5.66	41.2	2.05	23.8	12.0	0	缩孔

2.2 欠采样方法

直接对数据集里数量多的类别样例进行欠采样，去除一些类别多的样例使得各个类别的样例数据接近。如果随机丢弃样本，可能会丢失一些重要信息。

欠采样中最具代表性的算法是ENN（Edited Nearest Neighbor）算法。ENN算法通过两种策略删除不满足要求的样本点：Mode众数策略，计算每一个多数类样本的N个最近邻样本点，如果这N个最近邻样本点中多数样本点属于多数类，则删除该点，否则保留该点；All全部策略，计算每一个多数类样本的N个最近邻样本点，如果这N个最近邻样本点全部属于多数类，则删除该点，否则保留该点。

如果该点附近多数是或全部是多数类，则表明在决策程式判断时很有可能也会判断成多数类，同时表明该点所携带的绝大部分信息或者全部信息能够在近邻点中找到，所以该点携带了重复信息，删除对样本

的信息并没有太大的伤害。但是真实数据中这样的点是比较少的，只能有限删除真实数据中的少量点，并不能达到平衡数据的要求。

2.3 过采样方法

对数据集中的少数类样本进行过采样，就是将少数类样本相加，使分类问题中的每个类的样本数量接近。但是，过采样不能简单对少数类样本进行重复采样（即复制样本），否则会导致严重的过拟合。

过采样的代表性算法是SMOTE算法，SMOTE算法的生成过程为：

（1）对于少数类中的每个样本X_i，以欧氏距离为标准，计算其到少数类样本集中所有样本的距离，得到每个少数类样本X_i的k个最近邻；

（2）根据样本不平衡比设定一个抽样比n。对于每个少数类样本X_i，从其k个最近邻中随机选取几个样本，假设所选近邻为X_{zi}；

（3）对于每一个随机选出的近邻X_{zi}，分别与原样本按照如下的公式构建新的样本：

$$x_n=x_i+rand\ (0,1)\times(xzi-x_i) \quad (1)$$

SMOTE算法核心思想如图2所示。

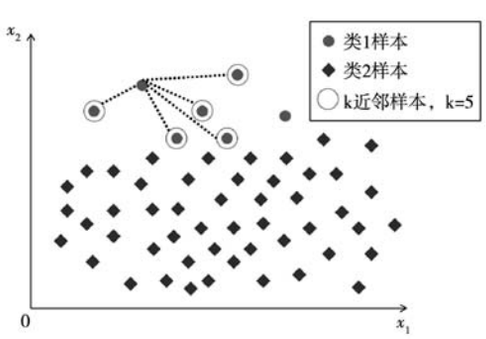


图2 SMOTE算法核心思想

Fig. 2 SMOTE algorithm core idea

3 SMOTE算法实现及效果分析

3.1 方法设计

数据集中每一条生产数据都是宝贵的，如果使用欠采样随机丢弃样本，可能会丢失一些重要信息。因此选择过采样方法增加各类缺陷数据使各类缺陷数据量与无缺陷数据量接近。方法设计图如图3所示。

3.2 SMOTE 算法实现

采用python语言实现SMOTE算法，具体步骤如下：

（1）导入需要的库，包括开放源代码软件库Tensorflow、科学计算库Numpy、读取数据文件库Xlrd、编写数据文件库Xlwt、随机数标准库Random、机器学习库Sklearn；

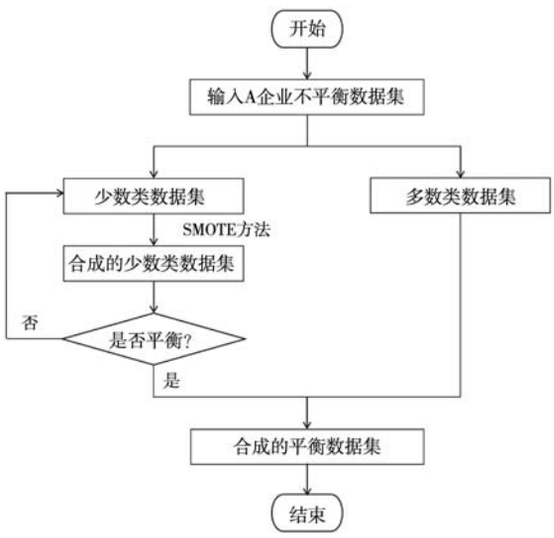


图3 方法设计图

Fig. 3 Method design diagram

（2）加载数据，利用读取数据文件库Xlrd读取Excel表中的数据，将归一化处理后的数据导入数组；

（3）创建SMOTE类，包括构造函数、寻找每一个样本在数据集X中的k近邻函数、分别根据其k近邻以及采样倍数生成新样本函数；

（4）导出合成数据集。

3.3 数据预处理效果分析

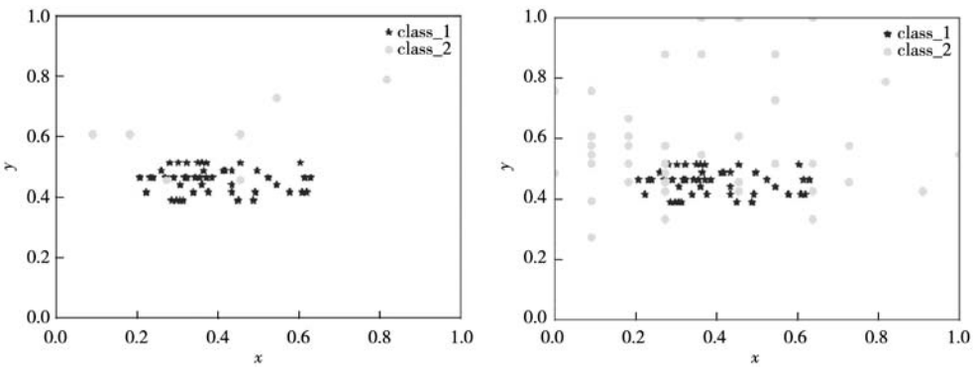
以样本属性中的浇注温度和混配土为例，使用SMOTE算法进行数据预处理的效果示意图如图4所示。对效果图进行简单分析，黄色圆点表示少数类样本数据，数据预处理后，可以清楚地看到，原来的类已经发展到令人满意的水平，并且结果类之间的差距并不大。

经过简单预处理的数据集，包含无缺陷数据5 848条，冷隔缺陷数据274条，气孔缺陷数据399条，砂眼缺陷数据359条，缩孔缺陷数据148条，属于不平衡数据。经过SMOTE算法进行数据预处理，得到总样本数据量30 862条，每一种类别，即无缺陷、冷隔缺陷、气孔缺陷、砂眼缺陷、缩孔缺陷的数据量各6 000条左右。至此，解决了数据不平衡的问题。数据预处理前后数据量对比如表2所示。

4 数据预处理对缺陷预测模型性能影响分析

4.1 缺陷预测模型性能衡量标准

模型的目标是通过样本的属性值来预测样本是



（a）数据预处理前　　（b）数据预处理后

图4　SMOTE算法效果示意图

Fig. 4 SMOTE algorithm effect diagram

表2　数据预处理前后数据量对比

Table 2 Comparison of data volume before and after data preprocessing

项目	无缺陷数据	冷隔缺陷	气孔缺陷	砂眼缺陷	缩孔缺陷
数据预处理前样本数量	5 848	274	399	359	148
数据预处理后样本数量	5 848	6 006	6 384	6 408	6 215

否具有缺陷，这是典型的分类问题。为了实现分类任务，使用Relu函数作为隐含层的激活函数，使用Softmax函数作为输出层的激活函数，使用准确率和交叉熵损失作为衡量模型优劣的因素。

Relu函数是一种线性整流函数，又称修正线性单元，是人工神经网络中常见的激活函数。它通常是指用斜率函数及其变体表示的非线性函数。一般来说，数学中的线性整流函数是指斜率函数，即：

$$f(x)=max\ (0,\ x) \tag{2}$$

在神经网络中，线性整流是神经元的激活函数，它定义了神经元经过线性变换之后的非线性输出，其中w为权值矩阵，b为偏置值。对于从上一级神经网络进入神经元的输入向量x，该神经元使用线性整流器激活函数将输出到下一层神经元或作为整个神经网络的输出。

$$max\ (0,\ w^Tx+b) \tag{3}$$

Softmax的含义是，它不唯一确定某个最大值，而是给每个输出分类结果分配一个概率值，代表属于每个类别的可能性。Softmax函数定义如下：

$$Softmax\ (z_i)=\frac{e^{z_i}}{\sum_{C=1}^{C}e^{z_c}} \tag{4}$$

其中z_i为第i个节点的输出值，C为输出节点数，即分类的类别数。多个类别的输出值可以通过Softmax函数转换为［0，1］和1范围内的概率分布。

准确率表示预测正确的所有样本占所有测试样本的比例：

$$准确率=\frac{正确分类的样本数}{总样本数}\times 100\% \tag{5}$$

在二分的情况下，即判断是否正确分类，模型最后需要预测的结果只有两种情况，对于每个类别我们的预测得到的概率为p和$1-p$，交叉熵损失函数的表达式为：

$$L=\frac{1}{N}\sum_i L_i=\frac{1}{N}\sum_i -\left[y_i\cdot\log(p_i)+(1-y_i)\cdot\log(1-p_i)\right] \tag{6}$$

其中N为样本数；y_i是样本i的标签，正类为1，负类为0；p_i是样本i预测为正类的概率。

4.2　数据预处理前缺陷预测模型效果分析

交叉熵损失随迭代次数的变化曲线如图5所示，交叉熵损失在5 000次的迭代过程中，整体呈现下降态

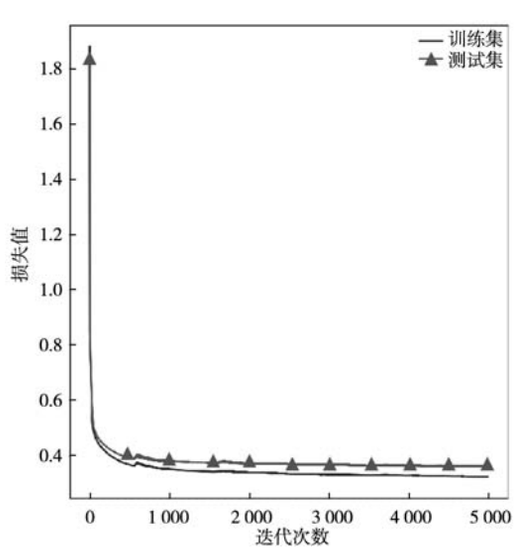


图5　交叉熵损失随迭代次数的变化曲线

Fig. 5 The curves of cross entropy loss variations with the numbers of iterations

势。但是在第600次迭代左右，出现了交叉熵损失增大的现象。我们需要得到最大似然估计，即模型的预测分布应该尽可能接近数据的实际分布，而交叉熵损失的增大，说明模型预测得到的分布与数据实际的分布误差增大。对于性能优良的神经网络预测模型，交叉熵损失随迭代次数的变化是严格递减的。

准确率随迭代次数的变化曲线如图6所示，准确率在5 000次的迭代过程中，整体呈现上升态势。但是在第800次至1 000次迭代中出现了剧烈的震荡，而准确率随迭代次数的变化应当是严格递增的，这说明采用本章节建立的缺陷预测模型不能准确预测铸件质量。

交叉熵损失和准确率随迭代次数的详细数值如表3所示。出现上述问题的原因，在于数据集的性质。获取的源数据，仅仅通过简单的预处理，并没有解决数据不平衡的问题。

4.3 数据预处理前后缺陷预测模型效果对比分析

进行数据预处理后，交叉熵损失随迭代次数的变化曲线如图7所示，与图5所示的数据预处理前模型效果相比，数据预处理后模型的交叉熵损失严格递减，且曲线光滑，未出现异常点，交叉熵损失值逐渐减小趋于0，说明基于数据预处理的缺陷预测模型预测出的结果与数据的实际分布情况大致相同，性能优良。

进行数据预处理后，准确率随迭代次数的变化曲线如图8所示，与图6所示的数据预处理前模型效果相比，数据预处理后模型准确率严格递增，且曲线光滑，未出现异常点，说明建立的缺陷预测模型能够进行准确预测，性能优良。经过多次预测模拟，最终预

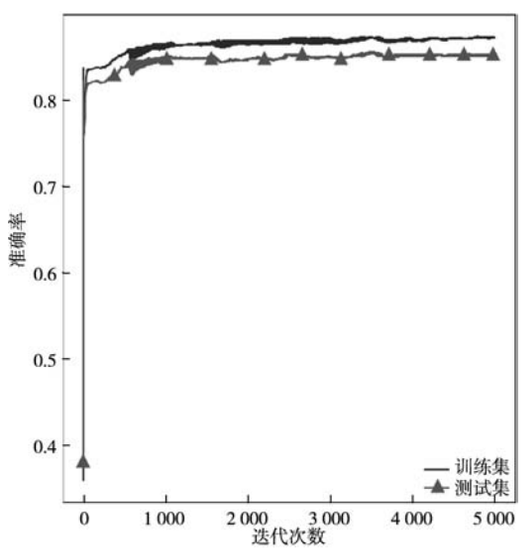


图6 准确率随迭代次数的变化曲线

Fig. 6 Accuracy curves with the number of iterations

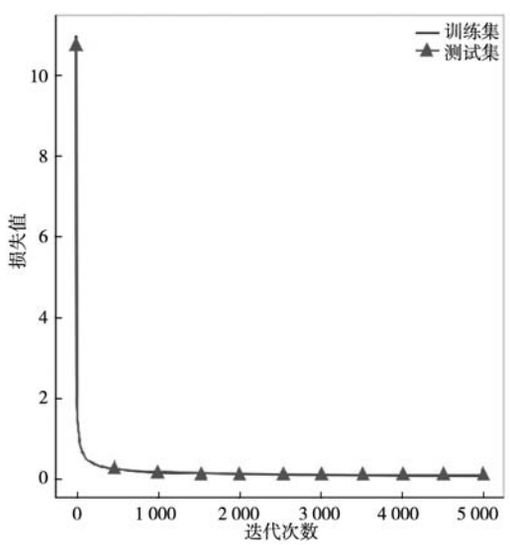


图7 交叉熵损失随迭代次数的变化曲线

Fig. 7 The curve of cross entropy loss with the number of iterations

表3 交叉熵损失和准确率随迭代次数变化的详细数值

Table 3 Detailed values of cross entropy loss and accuracy variations with the numbers of iterations

迭代次数	训练集准确率	训练集交叉熵损失	测试集准确率	测试集交叉熵损失
0	0.359 529	1.871 099	0.372 437	1.878 568
500	0.837 130	0.389 856	0.845 672	0.374 149
1 000	0.854 404	0.358 669	0.865 034	0.353 351
1 500	0.856 492	0.353 220	0.866 173	0.349 039
2 000	0.856 492	0.348 695	0.866 743	0.345 399
2 500	0.857 441	0.344 320	0.869 021	0.343 115
3 000	0.860 099	0.343 930	0.863 326	0.346 081
3 500	0.862 566	0.338 425	0.867 882	0.340 953
4 000	0.862 756	0.335 308	0.865 604	0.339 064
4 500	0.864 275	0.330 371	0.865 034	0.336 362
5 000	0.863 895	0.327 997	0.865 034	0.338 261

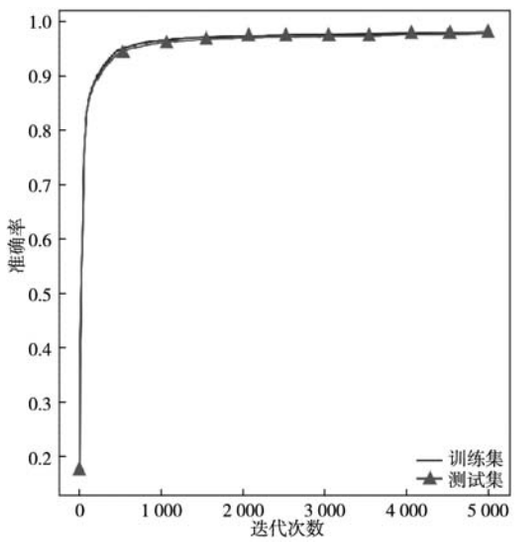


图8 准确率随迭代次数的变化曲线

Fig. 8 Accuracy curves with the numbers of iterations

测得到的训练集准确率约为97.99%，测试集准确率约为97.91%。

交叉熵损失和准确率随迭代次数的详细数值如表4所示。对表4中的数据进行分析，发现在整个迭代过程中，训练集和测试集的准确率随迭代次数严格递增，交叉熵损失随迭代次数严格递减，整个迭代过程未出现异常点。

5 结束语

针对实际生产过程采集到的复杂转向桥铸件的工艺数据中冷隔、气孔、砂眼、缩孔等缺陷类别的数据量严重不平衡、复杂铸件缺陷预测模型准确率不高的问题，开展了SMOTE数据预处理算法在砂型铸造复杂铸件缺陷预测中的应用研究，结果如下。

（1）获取了企业生产数据，并进行了数据预处理，分析了数据不平衡问题。

表4 交叉熵损失和准确率随迭代次数变化的详细数值

Table 4 Detailed values of cross entropy loss and accuracy variations with the numbers of iterations

迭代次数	训练集准确率	训练集交叉熵损失	测试集准确率	测试集交叉熵损失
0	0.208 100	10.912 581	0.204 566	10.967 226
500	0.947 350	0.221 916	0.944 301	0.225 462
1 000	0.964 400	0.154 114	0.961 978	0.159 136
1 500	0.970 300	0.126 612	0.967 962	0.132 265
2 000	0.972 550	0.111 444	0.971 276	0.116 983
2 500	0.974 900	0.101 415	0.973 209	0.106 645
3 000	0.975 700	0.093 852	0.974 406	0.098 821
3 500	0.976 250	0.087 857	0.975 143	0.092 671
4 000	0.977 200	0.082 917	0.976 155	0.087 609
4 500	0.978 450	0.078 798	0.978 089	0.083 401
5 000	0.979 900	0.075 302	0.979 101	0.079 816

（2）引入了SMOTE算法，科学扩充了不平衡数据集，创建了可用于模型训练的平衡数据集。

（3）数据预处理前后的模型预测准确率从86.50%提高至97.91%。

参考文献：

[1] XU Z，SHEN D，NIE T，et al. A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data [J]. Information Sciences，2021，572：574–589.

[2] FENG S，KEUNG J，YU X，et al. Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction [J]. Information and Software Technology，2021，139：106662.

[3] LIU S，JIANG H，WU Z，et al. Data synthesis using deep feature enhanced generative adversarial networks for rolling bearing imbalanced fault diagnosis [J]. Mechanical Systems and Signal Processing，2022，163：108139.

[4] KRISHNAN U，SANGAR P. A Rebalancing framework for classification of imbalanced medical appointment no-show data [J]. Journal of Data and Information Science，2021，6（1）：178–192.

[5] DUAN M，LIU D，CHEN X，et al. Self-balancing federated learning with global imbalanced data in mobile systems [J]. IEEE Transactions on Parallel and Distributed Systems，2020，32（1）：59–71.

[6] JING X Y，ZHANG X，ZHU X，et al. Multiset feature learning for highly imbalanced data classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2019，43（1）：139–156.
[7] WANG X，XU J，ZENG T，et al. Local distribution-based adaptive minority oversampling for imbalanced data classification [J]. Neurocomputing，2021，422：200–213.
[8] SYED M N，HASSAN M R，AHMAD I，et al. A novel linear classifier for class imbalance data arising in failure-prone air pressure systems [J]. IEEE Access，2020，9：4211–4222.
[9] ASIM Y，MALIK A K，RAZA B，et al. Predicting influential blogger′ s by a novel，hybrid and optimized case based reasoning approach with balanced random forest using imbalanced data [J]. IEEE Access，2020，9：6836–6854.
[10] FATHY Y，JABER M，BRINTRUP A. Learning with imbalanced data in smart manufacturing：a comparative analysis [J]. IEEE Access，2020，9：2734–2757.
[11] 潘徐政. 基于数据合成与机器学习的复杂铸件缺陷预测技术 [D]. 武汉：华中科技大学，2022.

Application of SMOTE Data Preprocessing Algorithm in Defect Prediction of Complex Castings in Sand Casting

PAN Xu-zheng[1], LIU Ying-hui[1], LI Wen[1], JI Xiao-yuan[1], YIN Ya-jun[1], WU Lai-fa[2], XIE Ming-guo[2], ZHOU Jian-xin[1]
(1. State Key Laboratory of Material Forming and Die Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei China; 2. Anhui Heli Co., Ltd., Hefei 230601, Anhui, China)

Abstract:
In view of the problems that the data amount of defects such as cold shut, blowhole, sand inclusion and shrinkage was seriously imbalance, and the accuracy of defect prediction model of complex casting was not high in the technological data of complex steering bridge casting collected during actual production processes, combined with the actual operating conditions of sand casting, the SMOTE (Synthetic Minority Oversampling Technique) data preprocessing algorithm was introduced to investigate its application in the defect predictions of complex castings manufactured by using of sand casting. According to the characteristics of the collected imbalance data set of complex castings, SMOTE data preprocessing algorithm was used to expand the imbalance data set scientifically and create a balance data set that can be used to train the defect prediction models of complex castings. The accuracies of the model predictions before and after data preprocessing increased from 86.50% to 97.91%.

Key words:
steering bridge casting; sand casting; imbalance data set; data preprocessing; SMOTE algorithm; defect prediction

（编辑：刘东辉，ldh@foundryworld.com）